Guidelines for the Management of Research Data at QUT

Originally prepared by P. Callan, J. Baker and L. De Vine

Edited by Philippa Frame & Sharron Stapleton

July 2018

Contents

* * * * *

Guidelines for the Management of Research Data at QUT

Originally Prepared by: Paula Callan, Janet Baker and Lance De Vine

Edited by: Philippa Frame, Sharron Stapleton

Last modified: 11th July 2018 (Updated link to Australian Code for the Responsible Conduct of Research, 2018)

Refer to MOPP D/2.8 for information about the QUT Management of Research Data Policy http://www.mopp.qut.edu.au/D/D_02_08.jsp.

For more information on the Guidelines and data management planning, see https://www.library.qut.edu.au/research/data/planning.jsp.

1 Introduction

These guidelines on the management of research data complement the QUT Policy for the Management of Research Data (MoPP D/2.8). The QUT Data Management Planning tool has been developed by QUT Library for use by researchers to assist in writing research data management plans.

Research data, in this context, refers to the information that is generated or collected to be used as primary sources in the production of original research results and would be required to validate or replicate research findings. Research data can be classified according to the processes used to gather or generate it:

- Experimental data: generated by lab equipment (e.g. gene sequences or chromatograms)
- Computational/Simulation data: generated from computation models – the actual model (and metadata about the model) may be more important than the output data (e.g. climate models, economic prediction models etc.)
- Observational data: recordings of specific phenomena at a specific time or location (e.g. seismic data, medical imaging, opinion polls, climate data etc.)
- Derived data, produced via the processing or combining of other data (e.g. data mining, compiled databases etc.)
- Canonical data: data extracted from reference datasets (e.g. GenBank, HILDA etc.).

Why Effective Management of Research Data is Important

Research data is a valuable product of the research process. Some data would be impossible to collect again; for example recordings of a specific seismic event. Good practices will ensure that researchers are able to meet any obligations related to data retention by guarding against the catastrophe of data loss. The effective management of research data also underpins research integrity by enabling the tracking of data from collection to results. This increases the repeatability of the research should it ever be necessary to defend a knowledge claim. The 'Australian Code for the Responsible Conduct of Research, 2018' which outlines high-level principles, responsibilities and expectations, indicates that researchers have a responsibility to manage research data effectively (https://www.nhmrc.gov.au/guidelines-publications/r41). Finally, the practices described in these guidelines will increase the usefulness of the data to other researchers if the data is able to be shared or published.

2 Research Data

2.1 Responsibility for Research Data

QUT's Code of Conduct for Research notes that "all research data, including primary materials, are considered to be University records and must be stored, disposed of or transferred in accordance with the QUT Records management policy (F/6.1). So the data collected as part of research studies are considered University records.

The QUT Management of research data policy (D/2.8) provides some useful insights into data management. Research data collected or generated as a part of research is defined as "data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or another research output is based. Data may be numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media."

All QUT researchers are responsible for maintaining records of research data and primary materials, and ensuring that these records and the research data and materials are securely stored. Researchers must make the data available to other researchers via open or negotiated access, as appropriate and in accordance with the requirements of research funding bodies such as the Australian Research Council (ARC) and the National Health and Medical Research Council (NHMRC) (see NHMRC revised policy on the dissemination of research findings) for as long as there is interest in the data."

2.2 Intellectual Property

Intellectual Property includes the following outcomes or products from research:

- patents for new or improved products or processes;
- trade marks for letters, words, phrases, sounds, smells, shapes, logos, pictures, aspects of packaging or a combination of these, to distinguish the goods and services of one trader from those of another;
- designs for the shape or appearance of manufactured goods;
- copyright for original material in literary, artistic, dramatic or musical works, films, broadcasts, multimedia and computer programs;
- circuit layout rights for the three-dimensional configuration of electronic circuits in integrated circuit products or layout designs;
- plant breeders' rights for new plant varieties;
- confidentiality / trade secrets including know-how and other confidential or proprietary information; and
- inventions.

2.2.1 Research Students

In accordance with general law principles, students personally own the intellectual property that they generate. However, in order to participate, or continue to participate, in certain projects students may need to assign to QUT their interest in the intellectual property, in particular, where: the project is externally funded, or the project has clear commercialisation potential or objectives. It is intended that requests for students to assign their intellectual property will be done only where necessary to enable QUT to meet its legal obligations to industry partners, or to achieve the objective of deriving appropriate benefits from innovation at QUT by commercialising intellectual property.

2.2.2 QUT staff

QUT owns the intellectual property generated by staff in the course of their employment. This means that QUT owns the copyright in research data that was generated or compiled by staff in the course of any normal duties covered by their contract of employment. However, the University assigns the right to publish "scholarly works" to the creators of those works subject to the conditions set out in chapter D/3.1.5. of the Manual of Policies and Procedures. That is, the assignment is subject to a non-exclusive licence in favour of QUT to allow QUT to use that work for teaching, research and commercialisation purposes and to reproduce and communicate that work online via QUT's open access digital repository

(refer to the University's Intellectual Property Policy, D/3.1. QUT retains all other rights in such scholarly works and all rights in works other than scholarly works.

Definition: "scholarly works" **means any article (refereed or un-refereed), book or book chapter(s), manual, musical composition or creative writing or any digital or electronic version of these that contains material written by staff or a student, but does not include work that is a computer program, teaching materials, or administration material or work produced at the request of a third party under contract with QUT.**

The rights retained by QUT in relation to scholarly works do not preclude the staff member from giving a similar **non-exclusive** 'right to publish' to other parties (for example, journal publishers). However, any granting of **exclusive** rights to a publisher would require the University to agree to waive its rights. Refer to MOPP D/3.1.5. http://www.mopp.qut.edu.au/D/D_03_01.jsp#D_03_01.05.mdoc.

2.2.3 Collaboration agreements

Collaboration agreements with external organisations may have an impact on ownership of the copyright in research data. At the commencement of any collaborative research project, parties should reach an agreement on matters of ownership. This agreement, which should be in writing, should cover ownership of any intellectual property that is produced, sharing of any commercial returns and ownership of copyrighted material. It should also cover various issues relating to the management of the research data including roles and responsibilities (who will do what), confidentiality, privacy, access, authorship, and dissemination of results. This agreement may take the form of a legal contract or a research management plan that has been signed by a representative from each organisation. An agreement will help resolve any conflicts over ownership that may arise later in the project. Generally, the University reserves the right to keep copies of data generated as a result of collaborative research.

Note: Academic staff should consult the Office of Commercial Services in relation to negotiation of research contracts with external parties generally or with the Office of Research in relation to negotiation of research contracts with external parties that include a competitive funding grant. Please note that QUT signing delegations require any contract with a value of greater than $10,000.00 to be executed by an officer within the Division of Research and Commercialisation or by the Vice-Chancellor.

For more information, see: https://www.qut.edu.au/research/scholarships-grants-and-funding/competitive-grant-funding QUT staff can access further information about competitive grant funding on the Digital Workplace: https://qutvirtual4.qut.edu.au/group/staff/research/competitive-grant-funding

3 Storage and Backup

Storage of Digital Data

Researchers must ensure that all research data is stored securely, in durable formats, securely from the commencement of the research project. Storage arrangements need to cover the statutory retention period. Refer to Section 7 of these guidelines for detailed information about statutory retention periods.

Things to consider when selecting where to store digital research data include:

- How long does the data need to be stored?
- Is security an issue?
- Could the hardware, software and media fail or become obsolete within the timeframe?
- Would the impact of such a failure be disastrous?
- Is support for the hardware, software or media available?
- Will others need to access the data in the future?

A number of QUT digital research data storage options exist to help QUT researchers store research data (see Appendix II). You can view an extended list of digital research data storage options under 'Data storage' on the Library website (https://www.library.qut.edu.au/research/data/storage.jsp).

3.1.1 Removable Media

Removable media such as USB drives, memory cards, CDs, DVD are convenient and affordable but, while they are suitable for working copies of data, they are not suitable for master copies or for long-term storage. Their portability means they can easily be lost. You may need multiple disks (or USB drives) making it difficult to retrieve specific files (especially if the documentation is poor). Removable media are also not very robust; the data can be damaged by magnetic fields, water and high temperatures.

If portable media are used for working copies of data, use only high-quality products and ensure that any confidential data is encrypted or password protected.

3.1.2 Desktop or Laptop Hard Drives

Storing master copies of research data files on the hard drive of individual desktop or laptop personal computers is not a good idea. This option may be convenient for working copies of data but the data should be backed up regularly to networked storage. Personal computers generally have a life expectancy of three years, but can fail at any time.

3.1.3 Networked Data Stores

Corporate networked storage is generally the best option for research data as the storage infrastructure is managed by IT experts. The data will be backed-up regularly, providing redundancy in the event that your computer crashes.

At QUT, ITS (Information Technology Services) has a consolidated platform providing 'tiered' storage for file servers at QUT. The platform is called ***estore***. It offers researchers large-scale networked storage for digital research data that can be accessed via a network drive. Files and documents saved on any of the network drives are automatically backed-up nightly in two physical locations.

The directory structure at QUT includes an "Archive" folder for each department. Inside this folder, the departmental network drive structure is mirrored. By moving a file to the Archive folder once you have finished using it regularly, it will be archived to ***estore***. To retrieve a file from archive storage you simply click on the filename, as you normally would, and it will be immediately recalled for you to open or edit.

Data that is stored on network drives may be accessible to a large number of people (e.g. a whole faculty or department). Network drives may also be configured, by appropriate IT staff, for use by a single user or group of users. Network drives provide a "medium" level security as some IT support staff will have access to the data as well as any authorised researchers. This type of security meets most general use-cases.

QUT's Research Data Storage service provides all QUT researchers (staff and Higher Degree Research students) with a secure data repository throughout the entire research data lifecycle. Raw research data can be acquired, project data can be worked on, published data can be archived. This storage suits the unique needs of research data.

Decisions about storage for highly confidential or highly sensitive research data should be made on a case-by-case basis, in consultation with IT support staff. The storage method and location may also need to be approved by a research ethics committee. See Section 6 of these guidelines for more information about privacy and confidentiality.

Statement of Authorship and Location of Data Form

Whenever a research-related publication is submitted to a journal a 'Statement of Authorship and Location of Data' form must be completed.

See MOPP D2.6 Code of Conduct for Research for access to the form: http://www.mopp.qut.edu.au/D/D_02_06.jsp.

The completed form should be retained by the school/centre or research unit and may be subsequently audited at the request of the Deputy Vice-Chancellor (Research and Commercialisation).

Note: The form requires the researcher to provide the location of the research data on which the publication is based. This is in case the publication stimulates questions that require reference to the data.

Non Digital Research Data

Data in non-digital formats (e.g. biological samples, analogue recordings) should be stored in secure facilities located in the School, Faculty, Institute or off-campus research facility. Refer to the Records Management section of the QUT Governance Services web page (http://www.governance.qut.edu.au/rms/) for more information about dealing with non-digital research records.

4 Data Formats

Durable Data Formats

When selecting formats in which to save research data, researchers should think about the long-term usability of the data and choose durable formats that will last for the lifetime of the project, plus the statutory retention period (if possible). Digital information is designed to be interpreted by software. Consequently, a risk to be considered is the possible obsolescence of the hardware and software required to access the data in the future. The required hardware and software may not be available towards the end of the required retention period.

Researchers should record information about which file format(s) they expect to use to capture and store their research data, and what software programs will be used to create and manipulate the data and retrieve the data (e.g. view, playback). If the software is unlikely to be widely available in the future, the software (and software manual) should be stored with the data. However, as operating systems change over time, no guarantees can be made as to long-term usability of any software.

A safer option is to use standard interchangeable formats that most software is capable of interpreting. Open formats, such as Rich Text Format (RTF) and Open Document Format (ODF), avoid 'lock-in' to proprietary formats which may not be accessible in the future. If the formats most suitable for the planned analyses are not open formats then, once the analyses have been completed, researchers should consider converting the data to standard interchangeable formats. When data is converted from one format to another, some data or internal metadata may be lost. Therefore, after any software conversion, the data should be checked for errors or changes.

Type of Data	**Durable Formats** (not comprehensive list)
Documentation	Rich Text Format (.rtf) PDF/A or PDF, HTML, Open Document Text (.odf) Plain text (.txt) XHMTL 1.0 Widely-used proprietary formats such as Word (.doc) and Excel (.xls)
Quantitative / tabular data	SPSS portable (.por), Stata, SAS , SPSS DDI XML comma-delimited (.csv), tab-delimited (.tab), MSExcel (.xls/.xlsx), MSAccess (.mdb/.accdb), dBase (.dbf) OpenDocument Spreadsheet (.ods)

Type of Data	**Durable Formats** (not comprehensive list)
Qualitative data (Text)	extensible mark-up language (.xml), Rich Text Format (.rtf), plain text ASCII (.txt), hypertext mark-up language (.html) NUD*IST, NVivo and ATLAS.ti
Digital video data	JPEG 2000
Digital image data	TIFF (version 6) uncompressed JPEG (.jpeg, .jpg) TIFF (other versions) Adobe Portable Document Format (PDF/A or PDF) raw image format (.RAW)
Digital audio data	Free lossless audio Codec (.flac) WAV (.wav) MPEG-1 Audio Layer 3 (.mp3) also OK
GIS and CAD (vector & raster)	ESRI Shapefile (.shp), GeoTIFF (geo-referenced TIFF) CAD data (.dwg) Binary formats of GIS and CAD also suitable

Table based on information disseminated by UK Data Archive (2008) http://www.data-archive.ac.uk/create-manage/format/formats-table.

5 Documentation and Metadata

Documentation

Good documentation adds value to research data as it ensures that the data will be easier to understand and the quality of the data will be easier to judge. Good records allow tracking from data collection to results, which supports reproduction and validation of the results. Documentation can also serve to guard against inappropriate use and accidental destruction.

Data documentation should provide contextual information for the data so that it can be understood in the future. Documentation requirements will vary depending on the discipline and type of research being conducted. Producing good documentation is easier if it is planned from the start of a project and considered throughout the lifecycle of the data.

Documentation should include:

- project aims and objectives (to provide context)
- catalogue of data collected
- description of lifecycle of key data elements (procedures for collection/creation, validation, transformation, processing, analysis, publication, archiving/destruction)
- description of instruments, calibrations etc.
- description of how data is structured (data model, coding schemes, controlled vocabularies etc.)
- details of any quality control processes
- confidentiality agreements and consent forms
- manuals, code books, procedure documents

Metadata

Metadata (data about data) is standardised information about a resource, presented in a structured format that is machine-readable and human-readable. Metadata can describe individual items or groups of items (individual files, images or datasets etc.). The items described by the metadata may be physical or digital. For example, a library catalogue includes metadata about books held in the library plus the electronic journals to which the library subscribes. The metadata helps the library to

manage its resources and assists users in the discovery and use of those resources. Likewise, metadata helps researchers to manage and re-use data after its creation

Ideally, as much metadata as possible should be gathered at the beginning of a research project, with ways devised to collect metadata (automatically if possible) throughout the life of the project. In general, the types of metadata collected will consist of:

- Descriptive metadata – includes descriptors for search and retrieval (bibliographic/descriptive terms) and references to publications pertaining to the data
- Rights metadata - information about ownership of data and any related publications
- Administrative metadata - management information including information about licences (permissions), confidentiality, access restrictions and timelines (e.g. release dates), preservation requirements
- Provenance metadata -information about the source, version tracking and transformations
- Technical metadata - information about file types, software used, size of file, contents of components of data such as variable names and descriptions or contributing performers of each track in audio recordings etc.
- Structural metadata – indicates how components of a set relate to one another. (e.g. a detailed list of all the tables in a database or the details of how one object relates to another (e.g. is earlier version of…)

A ‘metadata schema’ defines a set of terms that will be used to describe a resource and a set of rules that define the syntax or application and language (e.g. XML). Wherever possible, metadata should be created using an existing schema to assist in interoperability and the ability to share data.

Identifying Documentation and Metadata Requirements

The requirements will be shaped by short term and long term factors. In the short term you should consider the needs of the research project team. How will they retrieve and use the data? How will new research assistants know what to do? In the long term, if the plan involves depositing the data in a data repository or archive, you should consider the data documentation and metadata requirements of the relevant repository.

Things to consider:

- Responsibilities - who will be responsible for what?
- How will the metadata be stored? Some metadata can be stored within the digital object (e.g. the metadata in a digital image file) but often this is not the case. For external metadata, consider using a data repository. Repository software stores ‘digital objects’ which are made up of metadata plus one or more related files.
- How will the metadata be created? In some cases, metadata can be generated or extracted from digital files automatically. For example, a digital camera records the date, time, exposure setting, file format etc. In other cases, human effort will be required to create documentation and metadata. Software programmes sometime allow structured metadata (e.g. include title, author, organisation, subjects and keywords) to be added via "Properties".

Metadata Standards

If you are creating a database for personal use, you can use any terms you wish to ‘index’ the contents and, provided you are internally consistent, you will be able to retrieve sets of items that have been allocated the same ‘subject descriptor’. However, if you use a variety of descriptors for similar items, it will be much more difficult to retrieve them as a set.

Metadata standards and controlled vocabularies provide a means for standardising descriptions. Computers are then able to retrieve and merge metadata from multiple sources. Some standards are suitable for many different kinds of material and across disciplines others are more discipline-specific.

5.1.1 General Metadata Standards

Dublin Core

Dublin Core is widely used to describe digital materials such as video, sound, image, text, and composite media such as web pages. Each Dublin Core element is optional and may be repeated.

Dublin Core Metadata Initiative	
Simple Dublin Core Metadata Element Set (DCMES)	contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type
50 properties/terms in the DCMI specification	abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, contributor, coverage, created, creator, date, dateAccepted, dateCopyrighted, dateSubmitted, description, educationLevel, extent, format, hasFormat, hasPart, hasVersion, identifier, instructionalMethod, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, language, license, mediator, medium, modified, provenance, publisher, references, relation, replaces, requires, rights, rightsHolder, source, spatial, subject, tableOfContents, temporal, title, type, valid

Table based on information disseminated by DCMI http://dublincore.org/documents/dcmi-terms/

Dublin Core is used in a range of environments, not just for research projects. It is commonly used by institutional repositories, including QUT ePrints http://eprints.qut.edu.au.

Registry Interchange Format – Collections and Services (RIF-CS) Schema

The RIF-CS schema is used to describe 'collections' of research data, 'parties' (a person or group), 'activities' (event, project or program) and 'services' for the purpose of creating a metadata registry of research data collections. RIF-CS is used in QUT Research Data Finder (https://researchdatafinder.qut.edu.au/) as a guide for specifying, collecting and structuring metadata describing research data collections held locally. The metadata is subsequently be included in Research Data Australia; a registry of Australian research data collections.

For more information about RIF-CS, please refer to the ANDS websites http://services.ands.org.au/documentation/rifcs/guidelines/rif-cs.html and https://researchdata.ands.org.au/.

Discipline-specific Standards

In many disciplines, there are commonly-used standards for describing and sharing data within the discipline.

Humanities Data

- Functional Requirements for Bibliographic Records (FRBR) is used to describe bibliographic works and the relationship between a work and various manifestations or versions of the work (http://www.loc.gov/cds/downloads/FRBR.PDF)
- The Text Encoding Initiative (TEI) provides means of 'marking up' digital texts. The coding forms part of the digital text and is widely used in disciplines such as linguistics, literature and history (http://www.tei-c.org/index.xml)
- The Visual Resources Association Core (VRA) is used to describe original works of visual art and also images of them (http://www.loc.gov/standards/vracore/)
- Dublin Core is also widely used in the humanities (see section above) (http://dublincore.org/documents/dcmi-terms/)

Geospatial Data

- The Content Standard for Digital Geospatial Metadata (CSDGM), ISO 19115:2003 is suitable for geographical data as it identifies the extent, the quality, the spatial and temporal attributes, spatial reference, and distribution of digital geographic data (http://www.fgdc.gov/metadata/)
- ANZLIC is developing nationally-agreed (in both Australia and New Zealand) policies and guidelines aimed at achieving "best practice" in spatial data management. The ANZLIC

Metadata Profile was endorsed by the Spatial Information Council of Australia and New Zealand (formerly known as the Australia New Zealand Land Information Council) in 2007 (http://www.anzlic.gov.au/resources/anzlic_metadata_profile)

Social Sciences Data

- Data Documentation Initiative (DDI) is a standard for technical documentation describing social science data (http://www.icpsr.umich.edu/DDI/). The DDI element set can be mapped to the Dublin Core element set (http://www.ddialliance.org/resources/tools/dc)

Scientific Experimental Data

- CCLRC Scientific Metadata Model. The data model attempts to capture scientific activities at different levels: at the top level are 'Policies' which are enacted by initiating 'Programmes' which consist of one or more generic activities called 'Studies'. Each Study has one or more 'Investigations' which can be of different types (e.g. Measurement, Simulation, Experiment etc.) (http://www.dcc.ac.uk/resources/metadata-standards/csmd-cclrc-core-scientific-metadata-model)

There are many discipline-specific metadata standards available. Investigation of the commonly used standards in your discipline should be part of the data management planning process.

Multimedia Metadata Standards

The following standards have been developed to assist with the management of multimedia files:

- NISO Z39.87-2002 Technical Metadata for Digital Still Images. Documents image provenance and history (production metadata) and information that will help to ensure that the image data can be rendered accurately on output (to screen, print, or film) http://www.niso.org/apps/group_public/project/details.php?project_id=69.
- MPEG-7 is a standard for describing audio and video content. It combines content metadata (title, creator, rights, information about people, objects, and events represented in the multimedia file, etc.), as well as technical metadata about the file http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm.

Identifiers

An identifier is a label, reference number or name for a data object and forms a key part of your documentation and metadata. To be useful over the long-term, identifiers need to be:

- Unique - globally unique if possible, but at the very least unique within your particular systems and processes
- Persistent - the identifier should not change over time.

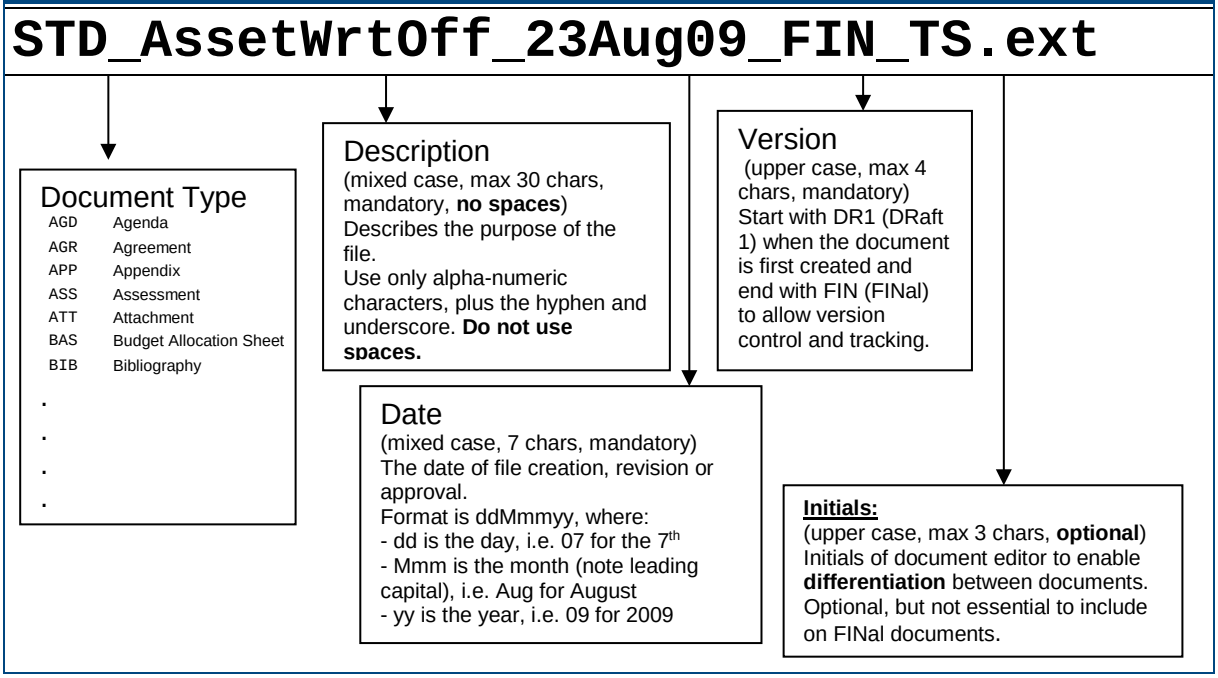
URLs are an example of an identifier as each URL is unique. However, if a published dataset is moved to a different location, the URL will return an HTTP 404 error (broken link) unless there is a commitment to maintain the persistence of the link via a redirection. A persistent identifier (PID) is an identifier that comes with a commitment that it will be kept up to date over a defined period of time. This can be achieved via a redirection implemented by the host institution or by the use of an identifier (which is embedded in a URL) supplied by a Persistent Identifier Scheme such as Digital Object Identifier (DOI) http://www.doi.org/, Persistent Uniform Resource Locators (PURL) http://purl.oclc.org or the Handle System http://www.handle.net/.

However, the data owner still needs to update the information if the location of the dataset changes. The persistence is a consequence of good practice rather than a technical solution.

For more information, refer to the Australian National Data Services' (ANDS) Guide on Persistent Identifiers, see: http://ands.org.au/guides/persistent-identifiers-awareness.html

File Organisation

In the data management planning stage, researchers should develop a protocol for naming the directory structure and digital files. For example, the top-level folder in the file directory should perhaps include the project title and year. The substructure could include folders for each run of an experiment, each version of a dataset, each person in the group, stages of the project or date periods. File names should be unique and based on a systematic naming convention. Use lower-case characters only – some computer operating systems are case-sensitive.Example of a file naming convention



For another example of a file naming convention, see: ARM Data Management and Documentation Plan, http://www.arm.gov/data/docs/plan.

Often, there are multiple versions of digital files so it is important to put in place version control procedures. For example:

- record version and status of a file, e.g. draft, interim, final, internal etc.
- record which changes are made to a file (by changing its filename)
- track the location of all files if stored in a variety of locations
- maintain single master files in a suitable format to remove version control problems associated with multiple working versions being developed in parallel.

A master file is a formalised and checked copy of the data (or other materials) at a specific stage of development as opposed to working versions of data. If a dataset is still active, a 'check-pointed' master file should be created for the data on which a publication was based. It is good practice to:

- assign responsibility for master files to specific members of the project team
- only the person with responsibility for the master file should have write access
- record any changes to master files
- create formal procedures for destroying master files (ideally, old master files should be retained).

Document Structures

Researchers should ensure that documents always contain titles, project name, author names, contact details, dates, and version information. Spreadsheets should always have unambiguous column and row labels. Each worksheet should be labelled and have a title row.

Data Dictionaries

A data dictionary helps to establish consistency by defining the attributes of data fields in a database or dataset, including the type and length of each data element. It is primarily a distillation of metadata about a database structure rather than the data itself.

Controlled Vocabularies

Controlled vocabularies mandate the use of predefined terms that have been pre-selected by the designer of the vocabulary. Examples include subject headings lists and taxonomies. Controlled vocabularies solve the problem of synonyms, where the same concept can be given different names (e.g. automobiles vs cars) and ambiguity, where the same word has multiple meanings. Some controlled vocabularies are very general (e.g. Library of Congress Subject Headings), while other are discipline–specific. Ontologies provide a controlled vocabulary, which can be used to model a domain. That is, an ontology describes the type of objects that exist, their properties and relationships.

If a suitable controlled vocabulary exists, it should be used as it adds shared meaning to the data. It also makes it more useful to other researchers in the same discipline.

Information Model

Documentation may need to include an information model that describes the data, the data sources and the relationship between them. A graphical/diagramming tool may be useful for creating this. Depending on the complexity of the data, a 'data workflow' diagram may also be constructed that shows the various processes involved in processing/transforming data complete with inputs and outputs.

6 Privacy and Confidentiality

Researchers are responsible for the ethical treatment of data. Research data which includes confidential or private information must be managed in accordance with any contractual or funding agreements.

The researcher must seek ethical clearance at the outset of the research project. Contact the QUT Office of Research Ethics and Integrity (http://www.orei.qut.edu.au/).

Where applicable, researchers should document:

- the nature of any private, sensitive or confidential information that may be collected
- non-disclosure agreements and any restrictions on use of the data
- consequences/penalties for breaches of confidentiality
- steps to be taken to safeguard privacy and confidentiality.

Privacy

Research data, particularly in health-related disciplines, may contain personal information about identified individuals. 'Personal' information is information that can be used to identify an individual. 'Sensitive' information includes health information or genetic information and is accorded higher levels of protection than other personal information.

As a statutory authority, QUT must comply with the requirements of the Information Privacy Act 2009, which is designed to provide safeguards for the handling of an individual's personal information in the public sector environment. The Act contains eleven Information Privacy Principles which provide guidance on the collection, storage, management, use and disclosure of personal information. http://www.governance.qut.edu.au/compliance/privacy/obligations.jsp

Researchers must obtain consent of participants to collect sensitive information. The process involves explaining the purposes, methods, risks and outcomes of the research including the extent to which the information will be disclosed to others. Where data has been 'de-identified' (for example by replacing names with numerical IDs) it can generally be disclosed to others provided re-use falls within the nature of the purposes covered by the consent.

In research involving humans, researchers should consult the National Statement on Ethical Conduct in Research Involving Humans and establish conditions of use for data gathered (http://nhmrc.gov.au/publications/synopses/e35syn.htm).

Confidentiality

A dataset or database may include information that is secret or confidential. The information can be protected from unauthorised access by use of technical mechanisms such as encryption, and passwords or legal mechanisms such as a confidentiality agreement. The actual or threatened disclosure of confidential information, which may include data that has not been made public, may result in legal liabilities.

A confidentiality agreement (also known as non-disclosure agreements) should be used where researchers wish to share confidential data on the understanding that it will not be further disclosed or used for purposes other than those covered by an agreement. An action for breach of confidence can be brought against someone who discloses or threatens to disclose confidential information to the detriment of those who wished to keep it confidential.

If a written agreement is in place, an action for breach of confidence can be brought against someone who discloses the confidential data to the detriment of those who wished to keep it confidential.

If you need to draft a confidentiality agreement for use with members of the research team, collaborators or for recipients of confidential data then contact the Office of Commercial Services. You will need to supply the relevant information about data owners and the data that is to be kept confidential.

Typically, the agreement will:

- identify the owners of rights in relation to the confidential data
- describe the data that is to be kept confidential
- oblige the person to whom the data is disclosed to maintain the confidentiality
- describe the extent of any permitted use
- define the consequences of any failure to comply with the confidentiality obligations

If the data include information contributed by or about indigenous peoples then indigenous knowledge systems and processes must be respected. This includes respecting Indigenous peoples' right to maintain the secrecy of Indigenous knowledge and practices. For information about the use of data related to studies involving indigenous Australians, consult the Australian Institute of Health and Welfare's National best practice guidelines for collecting Indigenous status in health data sets at http://www.aihw.gov.au/publication-detail/?id=6442468342.

See also: QUT | Office of Research Ethics and Integrity

7 Retention and Disposal of Research Data

Retention of Research Data

Research Data and research records must be retained for as long as required by legislation, statutory requirements, funding agency guidelines and contractual arrangements with research partners. Decisions about data retention and disposal should be documented, ideally in a data management plan, and stored with the data. The Queensland State Archives University Sector Retention and Disposal Schedule - https://www.forgov.qld.gov.au/schedules/university-sector-retention-and-disposal-schedule - specifies the retention periods for research data (See section 601.2/A50).

Retention Periods for Research Data *Research Data and records from research project that describe the research, permissions, contractual agreements and ownership of the data.	
Description (nature of the research)	**Retention Period**
Significant Research data which is of high public interest or significance to the discipline such that it has or will change a commonly held view or approach. Factors which may determine significance include projects which: • are controversial • are subject of extensive debate • arouse widespread scientific or other interest • have the potential to cause major adverse impacts on the environment, society or human health • involve eminent researchers • involve the use of major new or innovative techniques	**Permanent** (reference 601.2/C123)
Clinical Trials	**Retain for 15 years** after completion of clinical research/trial **AND 10 years** after last patient service provision or medico-legal action. (reference 601.3/C148)
Research data other (which does not result in a patent)	**Retain for 5 years** after last action. (reference 601.3/C150)
Research data other (which results in a patent)	**Retain for 7 years** after expiry of patent. (reference) 601.3/C149)

Disposal of Research Data

After the required retention period, the research data may be destroyed. If there is a possibility that the data includes sensitive or confidential information, secure data disposal methods should be adopted. Secure data disposal software or services should also be deployed when PCs that have been used in a research project are retired or recycled.

Disposal of non-digital data

- Complete the Records Disposal Application - Email Template. More information to be found at http://www.governance.qut.edu.au/rms/retention_disposal/index.jsp
- Physically destroy the records once approval is obtained from the Head of Section and Director, Governance Services
- Details of destruction and the date destroyed should be recorded via Email Template
- If the files are managed via a database, such as TRIM, the relevant record will need to be updated as 'destroyed' in the database.

Note: Some records, such as duplicates and drafts are classed as ephemeral records. These documents can be destroyed without completing a QUT Records Disposal Application.

8 Access and Re-use

There is considerable variation between disciplines in terms of inclination to share research data. Some disciplines, such as climatology and environmental science, have a long tradition of sharing data while

other are much less inclined to make their data accessible beyond the immediate research team. However, there are many benefits to be gained from making data as accessible as possible.

QUT is committed to the principle of 'Open Access' to the results of publicly funded scholarship and encourages its staff and students to generate and share knowledge that will provide social, cultural and economic benefits for QUT and the wider community. However, this must be balanced against the need to protect intellectual property in cases providing commercial opportunity or where considerable investment has been made by QUT or an industry partner. Consideration must be given to the prospect of commercialisation at the planning stage of a research project. Data may form part of an invention that is patentable and disclosure of the data could prevent a patent from being obtained so, while patentable research is not common, the issue should be given early consideration. If a patent is a possibility, data should only be released under confidentiality agreements until the patent has been obtained. Contact the Office of Commercial Services if you think the intellectual property has commercial potential.

The Australian Code for the Responsible Conduct of Research, 2018 suggests that, whenever possible, research data should be made available to other researchers (R22). Therefore, the default position should be to consider sharing the data unless there are compelling reasons why this is not appropriate. Even where the data cannot be made openly accessible, it may be that a subset of the data can be made accessible on request (in which case, metadata describing the dataset should be made openly accessible).

Repositories and Data Centres

Research data can be made accessible via a data repository (sometimes referred to as a data archive or data centre). Depositing data in a repository or archive is sometimes a condition attached to a grant or a requirement to be met when publishing a paper based on the data. Since 2008, the Australian Research Council (ARC) Funding Rules for Discovery Projects has encouraged researchers to deposit their data (as well as any publications) in an appropriate repository. If there is no intention to do this within six months, the grant recipient must provide an explanation in the final report. The National Health and Medical Research Council (NHMRC) has endorsed a similar policy (NHMRC's Policy on the Dissemination of Research Findings).

Data repositories may be institutional, where the data owners are affiliated with the institution, or discipline based, where the data has relevance to a specific discipline. The Australian Social Science Data Archive is an example of the latter. See Appendix 1 for a list of research data repositories and data centres.

Where research data can be made openly accessible via a data repository, the stable URL provided by a repository means the data can be cited. Alternatively, a repository can provide open access to just the metadata with negotiated access to the data itself. Depositing the data in a repository ensures that it can be stored safely for the relevant retention period (or longer). Conventions for directly citing datasets is still evolving but, in the meantime, it appears that the publications associated with openly accessible datasets are cited more frequently than similar publications in the same field where there is no access to the data.

Decisions need to be made, at the beginning of any research project, about whether or not data will be made available to others and, if so, on what basis?

Issues to consider include:

- Are there any legal constraints to data sharing (e.g. copyright, patents, privacy, confidentiality, or contracts)?
- Are there are conditions attached to the research grant which require the data to be deposited in a repository or archive?
- If the data can be made openly accessible:
- Which data is to be made available?
- When is the data to be made available (e.g. at the end of the project)?
- Which repository or archive would be most suitable?
- What re-use rights will attach to the data?
- How will the re-use rights be communicated?
- For data that cannot be made openly accessible:

- Who will have access (e.g. just the research team)?
- Is it possible to deposit metadata in a repository?
- Is access 'by negotiation' to de-identified data a possibility?

De-Identifying Research Data

Data about people should be de-identified before it is shared or published. For ethical and legal reasons, it should not be possible to identify particular individuals or organisations from the data.

Quantitative datasets can be de-identified by removing names and addresses or by reducing the level of precision (e.g. aggregate to a higher level). Special care is needed where access to multiple variables could accidentally disclose the identity of a person (e.g. workplace plus occupation and age). Also, geo-referenced data may disclose the identity of an organisation by disclosing an exact location.

For qualitative data, names could be replaced with numerical identifiers or pseudonyms. Retain intact version of the data for use by the research team. Document all the steps taken to de-identify the dataset. De-identifying video data can be difficult and costly. If possible obtain participant consent to share the data.

Access Agreements

In some cases, researchers may be willing to allow access to research data provided the persons requesting access are prepared to enter into an access agreement. The agreement allows the data owners to specify the terms of access. For example, they may specify that any profits will be shared. Access agreements for QUT materials are negotiated by the Division of Research and Commercialisation or qutbluebox and should be consulted in relation to any proposed access agreement. In consulting with the Division, the following information would be required:

- Description of the data;
- Identity of those who are permitted to access the data;
- The limits of the access rights (e.g. what cannot be done with the data)
- A provision for consequences for failure to comply with the agreement

8.1.1 Copyright Licensing

A copyright licence is permission, granted by the copyright owner, which allows another person to exercise one or more of the rights of the copyright owner including the right to reproduce the dataset, migrate files to a new format or communicate the dataset to the public. This has implications for data management as it means that the copyright owner must grant a licence (deposit licence) to a repository to enable that repository to store a dataset (as this involved making at least one copy) and (if applicable) to communicate that dataset to the public. If the dataset is openly accessible online via the repository, the end-user needs the copyright owner's permission to use the dataset. This licence is usually attached to the dataset in the repository so end-users can determine limits to use granted by the copyright owner.

8.1.2 Types of Licences

Licences can be exclusive or non-exclusive. An exclusive licence means that only the recipient (licensee) has the right to use the data in the manner covered by the licence – to the exclusion of the copyright owner. A non-exclusive licence means that the recipient (licensee) has the right to use the data in the manner covered by the licence – but not to the exclusion of the copyright owner who is free to use the data in the same manner and to grant non-exclusive licences to other licensees. Data owners generally grant non-exclusive licences to repositories and other end-users of their data.

Licences can also be:

- time limited (for example, a licence for one-off use or for a specified period of time)
- limited to geographical regions (for a example, only applies within Australia)
- limited to specific purposes (for example, non-commercial use only)
- contractual (this means that the licensee agrees to do something in return)
- non-contractual (this means that there is no reciprocal obligation but restrictions may still apply).

'Open Content' licences are non-contractual licences where the data owner expresses (via the visible licence) how the data can and cannot be used. This facilitates open access and re-use as it frees the end-user from having to contact the data owner to ask for permission. Example: Creative Commons (CC) Licences.

Creative Commons licences grant copyright permission to the world-at-large to use the licensed material, on certain conditions. A CC licence is a 'some rights reserved' licence. The most liberal CC licence reserves only the right to be attributed as the owner. Other CC licences reserve this right (attribution) and specify one or more conditions which apply to end users. For example, 'non-commercial' use only, 'no derivatives' allowed or must 'share-alike'. Where data from multiple CC licensed datasets are merged, it will be necessary to consider whether the data taken is a substantial part of the dataset (if not, the attribution requirement does not arise) and if it is, the attribution requirements under the CC licences will need to be managed. There are techniques for easily attributing multiple creators, e.g. by listing the names of all creators in an online list and using the list's URL as a short-form attribution notice. If a copyright owner is happy to relinquish their right to be attributed, they can include a note saying that no citation is required, and that the author consents to not being named. https://creativecommons.org/licenses/

Note: Copyright and intellectual property owned by QUT must be licensed through the Division of Research and Commercialisation or qutbluebox. Please consult with the Division of Research and Commercialisation in relation to any such licensing matters.

9 References, Sources and Contacts

References

Dublin Core Metadata Initiative (2008) DCMI metadata terms http://dublincore.org/documents/dcmi-terms/

Fitzgerald, A, Pappalardo, K. Austin, A. (2008) *Practical data management: a legal and policy guide*. OAK Law Project. Retrieved October 22, 2009 from http://eprints.qut.edu.au/14923/

National Health and Medical Research Council, the Australian Research Council and Universities Australia (2018) *Australian Code for the Responsible Conduct of Research, 2018.* Retrieved July 11, 2018 from https://www.nhmrc.gov.au/guidelines-publications/r41

Piwowar H.A., Day R.S., Fridsma D.B., 2007 *Sharing Detailed Research Data Is Associated with Increased Citation Rate*. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

Research Information Network (2008) *Stewardship of digital research data - principles and guidelines*. London: RIN. Retrieved October 22, 2009 from http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines

UKData Archive (2008) *Manage and Share Data* http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Sources

The following sources are acknowledged as sources of inspiration and information used in the preparation of these guidelines.

ANDS Guide - Copyright and data http://ands.org.au/guides/copyright-and-data-awareness.html

ANDS Guide. Persistent Identifiers http://ands.org.au/guides/persistent-identifiers-awareness.html

ARM Data management and documentation plan https://www.arm.gov/data/docs/plan

Australian Code for the Responsible Conduct of Research, 2018. https://www.nhmrc.gov.au/guidelines-publications/r41

Australian Research Council (ARC) Funding Rules for Discovery Projects http://www.arc.gov.au/funding-rules

DISC-UK (2009) Policy-making for research data in repositories: a guide. http://www.disc-uk.org/docs/guide.pdf

Inter-University Consortium for Political and Social Research (ICPSR) Guide to Scoial Science data Preparation and Archiving. http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/ack.html

JISC Digital media Advice http://www.jiscdigitalmedia.ac.uk/advice/

Monash University. Research data management guidelines http://www.researchdata.monash.edu.au/

National Health and Medical Research Council (NHMRC). Dissemination of Research Findings http://www.nhmrc.gov.au/grants/policy/dissemination.htm

University of Edinburgh. (2011). MANTRA: Research Data Management Training. http://datalib.edina.ac.uk/mantra/

Research Information Network (2008) *Stewardship of digital research data - principles and guidelines*. London: RIN. Retrieved October 22, 2009 from http://www.rin.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines

Contacts

Paula Callan
Scholarly Communications Librarian
Email: p.callan@qut.edu.au
Phone: 3138 8336

Stephanie Bradbury
Research Support Coordinator
Email: s.bradbury@qut,edu.au
Phone: 3138 80618

Philippa Frame
Research Data Librarian
Email: library.research@qut.edu.au
Phone: 3138 0624

Nerida Quatermass
University Copyright Officer
Email: qut.copyright@qut.edu.au
Phone: 3138 4616

APPENDIX I

Research data repositories and data centres (selection)

Australian Data Repositories	
Australian Data Archive	http://assda.anu.edu.au/
Aboriginal & Torres Strait Islander Data Archive	http://www.atsida.edu.au/
BlueNet: Australian Marine Science Data Network	http://www.bluenet.org.au/
CSIRO Marine and Atmospheric Research Data Centre	http://www.cmar.csiro.au/datacentre/
Geoscience Australia	http://www.ga.gov.au/products-services/data-applications.html
Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)	http://www.paradisec.org.au/
International Data Repositories and Data Centres	
Archaeology Data Service	http://ads.ahds.ac.uk/
Biomedical Informatics Research Network (BIRN) Data Repository	http://www.birncommunity.org/resources/data/
British Household Panel Survey	https://www.iser.essex.ac.uk/bhps
Cambridge Structural database	http://www.ccdc.cam.ac.uk/products/csd/
Comparative Study of Electoral Systems	http://www.cses.org/
EarthChem	http://www.earthchem.org/
EDINA / University Data Library	http://edina.ac.uk/
Entrez (includes GenBank)	http://www.ncbi.nlm.nih.gov/Database/index.html
Environmental Information Data Centre	http://www.ceh.ac.uk/sci_programmes/env_info.html
History Data Service	http://hds.essex.ac.uk/
Social Science Data Archive	http://www.sscnet.ucla.edu/issr/da/
Inter-University Consortium for Political and Social Research (ICPSR)	http://www.icpsr.umich.edu/icpsrweb/ICPSR/
International Spectroscopic Data Bank	http://www.is-db.org/
Knowledge Network for Biocomplexity	http://knb.ecoinformatics.org/index.jsp
National Centre for Social Research	http://www.natcen.ac.uk/
National Geosciences Data Centre	http://www.bgs.ac.uk/services/ngdc/
National Digital Archive of Datasets	http://www.ndad.nationalarchives.gov.uk/
NERC Earth Observation Data Service	http://www.neodc.rl.ac.uk/
NZ Social Science Data Service	http://www.nzssds.org.nz/links
Petrological Database of the Ocean Floor	http://www.petdb.org/
System for Earth Sample Registration	http://www.geosamples.org/
TreeBASE	http://treebase.org/treebase-web/search/studySearch.html
UK Data Archive	http://www.data-archive.ac.uk/
Visual Arts Data Service	http://www.vads.ac.uk/
World Data System	http://www.icsu-wds.org/
World Wide Protein Data Bank (PDB)	http://www.wwpdb.org/